

# VIslual TRAnslator: Linking Perceptions and Natural Language Descriptions

Gerd Herzog, Peter Wazinski

SFB 314, Project VITRA  
Universität des Saarlandes  
D-66041 Saarbrücken  
vitra@cs.uni-sb.de

## Abstract

Despite the fact that image understanding and natural language processing constitute two major areas of AI, there have only been a few attempts towards the integration of computer vision and the generation of natural language expressions for the description of image sequences. In this contribution we will report on practical experience gained in the project VITRA (VIslual TRAnslator) concerning the design and construction of integrated knowledge-based systems capable of translating visual information into natural language descriptions. In VITRA different domains, like traffic scenes and short sequences from soccer matches, have been investigated.

Our approach towards *simultaneous* scene description emphasizes concurrent image sequence evaluation and natural language processing, carried out on an *incremental* basis, an important prerequisite for real-time performance. One major achievement of our cooperation with the vision group at the Fraunhofer Institute (IITB, Karlsruhe) is the automatic generation of natural language descriptions for recognized trajectories of objects in real world image sequences. In this survey, the different processes pertaining to high-level scene analysis and natural language generation will be discussed.

**This article first appeared in: Artificial Intelligence Review, 8 (2/3), pp. 175–187, 1994.**

**It has been reprinted in: P. Mc Kevitt (ed.), Integration of Natural Language and Vision Processing: Computational Models and Systems, Volume 1, pp. 83–95. Dordrecht: Kluwer, 1995.**

# 1 Introduction

Computer vision and natural language processing constitute two major areas of research within AI, but have generally been studied independently of each other. There have been only a few attempts towards the integration of image understanding and the generation of natural language descriptions for real world image sequences.

The relationship between natural language and visual perception forms the research background for the VITRA project (cf. Herzog et al. [1993b]), which is concerned with the development of knowledge-based systems for natural language access to visual information. According to [Wahlster, 1989, p. 479], two main goals are pursued in this research field:

1. “The complex information processing of humans underlying the interaction of natural language production and visual perception is to be described and explained exactly by means of the tools of computer science.”
2. “The natural language description of images is to provide the user with an easier access to, and a better understanding of, the results of an image understanding system.”

It is characteristic of AI research, that, apart from the cognitive science perspective (1), an application-oriented objective is also pursued (2). From this engineering perspective, the systems envisaged here could serve such practical purposes as handling the vast amount of visual data accumulating, for example, in medical technology (Tsotsos [1985], Niemann et al. [1985]), remote sensing (Bajcsy et al. [1985]), and traffic control (Wahlster et al. [1983], Neumann [1989], Walter et al. [1988], Koller et al. [1992b], Kollnig and Nagel [1993]).

The main task of computer vision is the construction of a symbolic scene representation from (a sequence of) images. In the case of image sequence analysis, the focus lies on the detection and interpretation of changes which are caused by motion. The intended output of a vision system is an explicit, meaningful description of visible objects. One goal of approaches towards the integration of computer vision and natural language processing is to extend the scope of scene analysis beyond the level of object recognition. Natural language access to vision systems requires processes which lead to conceptual units of a higher level of abstraction. These processes include the explicit description of spatial configurations by means of spatial relations, the interpretation of object movements, and even the automatic recognition of presumed goals and plans of the observed agents. Based upon such high-level scene analysis, natural language image descriptions have the advantage, that they allow variation of how condensed a description of visual data will be according to application-specific demands.

In VITRA, different domains of discourse and communicative situations are examined with respect to natural language access to visual information. Scenarios under investigation include:

- Answering questions about observations in traffic scenes (cf. Schirra et al. [1987])

- Generating running reports for short sections of soccer games (cf. André et al. [1988], Herzog et al. [1989])
- Describing routes based on a 3-dimensional model of the University Campus Saarbrücken (cf. Herzog et al. [1993a], Maaß et al. [1993])
- Communicating with an autonomous mobile robot (cf. Lüth et al. [1994])

In this survey, we will concentrate on our joint work with the vision group at the Fraunhofer Institute (IITB, Karlsruhe) regarding the automatic interpretation of dynamic imagery.

## 2 The Visual Translator

The task of the vision group at the IITB is to recognize and to track moving objects within real world image sequences. Information concerning mobile objects and their locations over time together with knowledge about the stationary background constitutes the so-called *geometrical scene description*. In Neumann [1989] this intermediate geometrical representation, enriched with additional world knowledge about the objects, has been proposed as an *idealized* interface between a vision component and a natural language system.



Figure 1: Three frames from the soccer domain

First results had been obtained in the investigation of traffic scenes and short sequences from soccer matches (cf. Fig. 1). Apart from the trajectory data supplied by the vision component ACTIONS (Sung and Zimmermann [1986], Sung [1988]), synthetic data have been studied in VITRA as well (c.f. Herzog [1986]). Since an automatic classification and identification of objects is not possible with ACTIONS, object candidates are interactively assigned to previously known players and the ball. The more recent XTRACK system (Koller [1992], Koller et al. [1992a]) accomplishes the automatic model-based recognition, tracking, and classification of vehicles in traffic scenes.

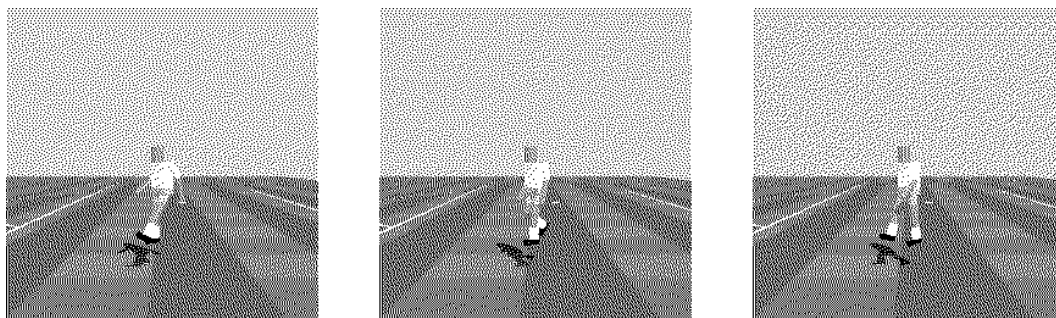


Figure 2: Geometric model of a human body

Research described in Rohr [1994] concentrates on the model-based 3D-reconstruction of *non-rigid* bodies. A cylindrical representation and a kinematic model of human walking, which is based on medical data, is utilized for the incremental recognition of pedestrians and their exact state of motion. This approach for the geometric modeling of an articulated body has been adopted in VITRA in order to represent the players in the soccer domain (cf. Herzog [1992b]). In Fig. 2 different movement states of the walking cycle are shown.

The goal of our joint efforts at combining a vision system and a natural language access system is the automatic *simultaneous* description of dynamic imagery. Thus, the various processing steps from raw images to natural language utterances must be carried out on an *incremental* basis. Fig. 3 shows how these processes are organized into a cascade within the VITRA system.

An image sequence, i.e., a sequence of digitized video frames, forms the input for the processes on the sensory level. Based on the visual raw data, the image analysis component constructs a geometrical representation of the scene, stating the locations of the visible objects at consecutive points in time. The contents of the geometrical scene description, which is constructed incrementally, as new visual data arrive, are further interpreted by the processes on the cognitive level. This high-level scene analysis extracts spatial relations, interesting motion events, as well as presumed intentions, plans, and plan interactions of the observed agents. These conceptual structures bridge the gap between visual data and natural language concepts, such as spatial prepositions, motion verbs, temporal adverbs and purposive or causal clauses. They are passed on to the processes on the linguistic level which transform them into natural language utterances. In terms of reference semantics, explicit links between sensory data and natural language expressions are established.

VITRA provides a running report of the scene it is watching for a listener who cannot see the scene her/himself, but who is assumed to have prior knowledge about its static properties. In order to generate communicatively adequate descriptions, the system must anticipate the visual conceptualizations that the system's utterance elicits in the listener's mind (cf. Neumann [1989], Wahlster [1989]). A peculiarity in VI-

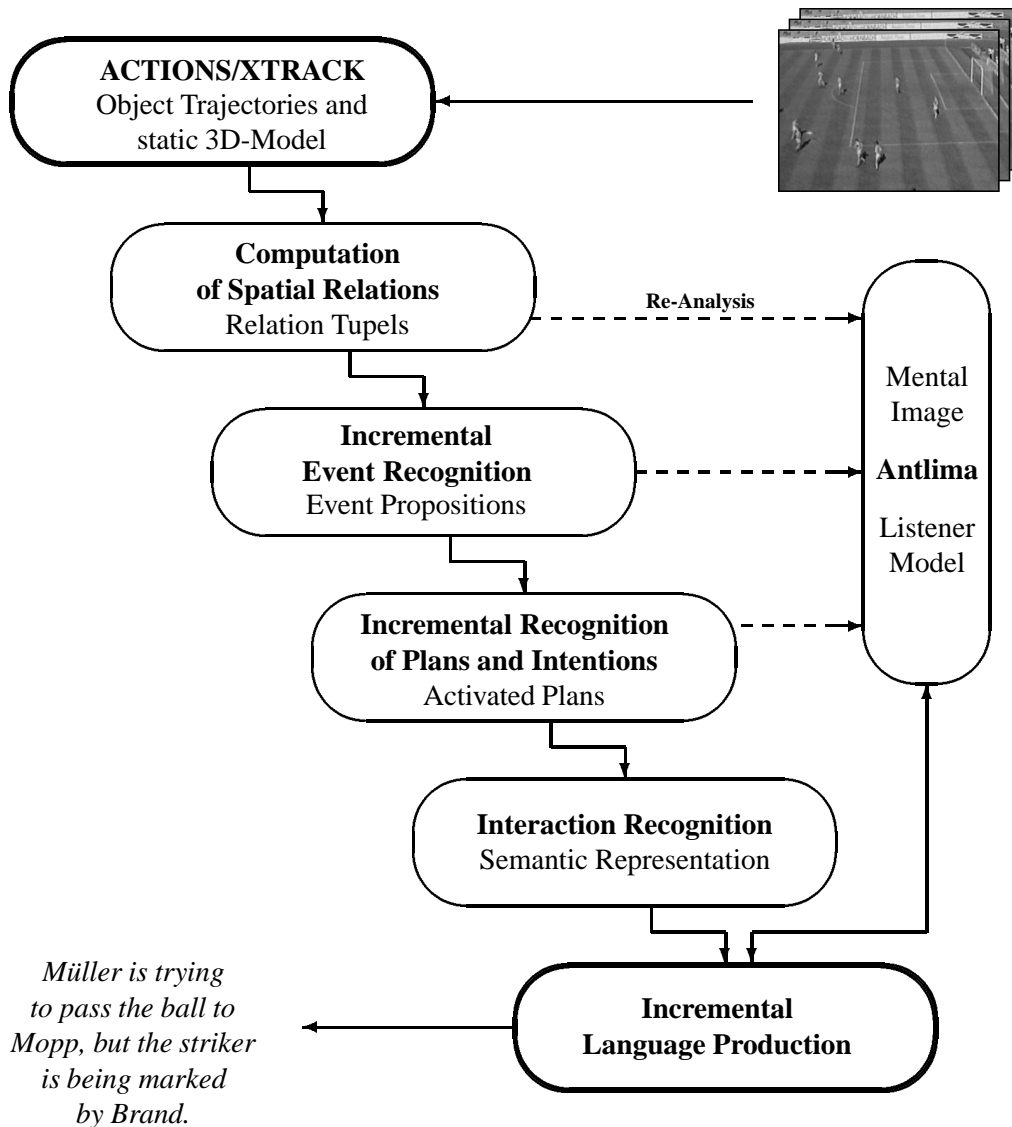


Figure 3: Cascaded processing in VITRA

TRA is the existence of such a listener model. Depending on the current text plan the component ANTLIMA is able to construct a mental image corresponding to the (assumed) imagination of the listener. This mental image is re-analyzed and compared with the system's visual information. Possible discrepancies may lead to changes in the preliminary text plan.

### 3 Incremental high-level scene analysis

Natural language access systems like HAM-ANS (Wahlster et al. [1983]) and NAOS (Neumann and Novak [1986]) concentrate on an *a posteriori* analysis. Low level vision processing considers the entire image sequence for the recognition and cueing of moving objects; motion analysis happens afterwards, based on complete trajectories. Since only information about a past scene can be provided, these systems generate *retrospective* scene descriptions. In VITRA we favour an *incremental analysis*. Input data is supplied and processed simultaneously as the scene progresses. Information about the present scene is provided and immediate system reactions (like motor actions of a robot, *simultaneous* natural language utterances) are possible.

#### 3.1 Interpreting spatial relations and object movements

The definition and representation of the semantics of spatial relations is an essential condition for the synthesis of spatial reference expressions in natural language. The computation and evaluation of spatial relations in VITRA is based on a multilevel semantic model, that clearly distinguishes between context specific conceptual knowledge and the basic meaning of a spatial relation (cf. Gapp [1994]).

The detailed geometric knowledge, grounded in visual perception, can be exploited for the definition of a reference semantics, that does not assign simple truth values to spatial predications, but instead introduces a measure of degrees of applicability that expresses the extent to which a spatial relation is applicable (cf. André et al. [1989]). Since different degrees of applicability can be expressed by linguistic hedges, such as *'directly'* or *'more or less'*, more exact scene descriptions are possible. Furthermore, if an object configuration can be described by several spatial predications, the degree of applicability is used to select the most appropriate reference object(s) and relation(s) for verbalization.

In the context of the VITRA project, different classes of spatial relations have been examined in more detail. Wazinski [1993a] and Wazinski [1993b] are concerned with topological relations. Orientation-dependent relations are treated in André et al. [1987a] and André et al. [1989]. Since the frame of reference is explicitly taken into account, the system is able to cope with the *intrinsic*, *extrinsic*, and *deictic* use of directional prepositions (cf. Retz-Schmidt [1988]). Recently, the algorithms developed so far have been generalized for 3-dimensional geometric representations (cf. Gapp [1993], Gapp [1994]).

If a real-world image sequence is to be described simultaneously as it is perceived, one has to talk about object motions even while they are currently happening and not yet completed. Thus, motion events have to be recognized stepwise as they progress and event instances must be made available for further processing from the moment they are noticed first. Consider the examples given in Fig. 4, where a white station wagon is passing a pick-up truck, and in Fig. 1, where a player is transferring the ball to a team mate.



Figure 4: A passing event in a traffic scene

Since the distinction between events that have and those that have not occurred is insufficient, we have introduced the additional predicates *start*, *proceed*, and *stop* which can be used to characterize the progression of an event (cf. André et al. [1988]). Labeled directed graphs with typed edges, so called *course diagrams*, are used to model the prototypical progression of an event. The recognition of an occurrence can be thought of as traversing the course diagram, where the edge types are used for the definition of our basic event predicates. Course diagrams rely on a discrete model of time, which is induced by the underlying image sequence. They allow incremental event recognition, since exactly one edge per unit of time is traversed. Using constraint-based temporal reasoning, the course diagrams are constructed automatically from interval-based concept definitions (cf. Herzog [1992a]).

The event concepts are organized into an abstraction hierarchy, based on specialization (e.g., walking is a moving) and temporal decomposition (e.g., passing consists of swing-out, drive-beside, and swing-into-line). This conceptual hierarchy can be utilized in the language production process in order to guide the selection of the relevant propositions.

### 3.2 Recognizing intentions, interactions, and causes of plan failures

Human observers do not only pay attention to the spatio-temporal aspects of motion. They also make assumptions about intentional entities underlying the behaviour of other people (e.g., player A does not simply *approach* player B, but he *tackles* him).

One criterion for the choice of soccer as a domain of discourse in VITRA was the fact that the influence of the agents assumed intentions on the description is particularly obvious here. Given the position of players, their team membership and the distribution of roles in standard situations, stereotypical intentions can be assumed for each situation. As described in Retz-Schmidt [1991] and Retz-Schmidt [1992], the VITRA system is able to incrementally recognize intentions of and interactions between the agents as well as the causes of possible plan failures.

Partially instantiated plan hypotheses taken from a hierarchically organized plan library are successively instantiated according to the incrementally recognized events. The leaves of the plan hierarchy represent observable events and spatial relations. An inner node corresponds to an abstract action. An edge, that connects two nodes either represents a decomposition or a specialization relation. In addition, a node also contains information about necessary preconditions of the action it represents as well as information about its intended effect.

In a continually changing domain it would be computationally intractable to keep track of all agents that occur in the scene. Therefore, domain specific focussing heuristics are applied in order to reduce the number of agents whose actions have to be observed. In the soccer domain, for example, the system would focus on the agents that are near the goal or the player who has the ball.

Knowledge about the cooperative (e.g., `double-pass`) and antagonistic behaviour (e.g., `offside-trap`) of the players is represented in the interaction library. A successful plan triggers the activation of a corresponding interaction schema. Similar to the plan recognition process this interaction schema has to be fully instantiated before the particular interaction is recognized.

There are several possibilities for a plan failure that can be detected with respect to the underlying plan and interaction recognition component: (i) An agent might assume a precondition for a plan that is not given, (ii) an antagonistic plan can lead to a plan failure, or (iii) in case of an cooperative interaction the partner fails.

## 4 Simultaneous natural language description

Since an image sequence is not described *a posteriori* but rather as it progresses, the complete course of the scene is unknown at the moment of text generation. In addition, temporal aspects such as the time required for text generation and decoding time of the listener or reader have to be considered for the coordination of perception and language production. These peculiarities of the conversational setting lead to important consequences for the planning and realization of natural language utterances (cf. André et al. [1987b]). As the description should concentrate on what is currently happening, it is necessary to start talking about motion events and actions while they are still in progress and not yet completely recognized. In this case encoding has to start before the contents of an utterance have been planned in full detail. Other characteristics of simultaneous reporting besides incremental generation of utterances need to be dealt with. The description often lags behind with respect to the occurrences in the scene and unexpected topic shifts occur very frequently.

Language generation in VITRA includes processes that handle the selection, linearization and verbalization of propositions (cf. André et al. [1988]). The listener model provides an imagination component, in order to anticipate the listener's visual conceptualizations of the described scene.

## 4.1 Selection and linearization of propositions

As the time-varying scene has to be described continuously, language generation underlies strong temporal restrictions. Hence, the system cannot talk about all events and actions which have been recognized, but instead it has to decide which propositions should be verbalized in order to enable the listener to follow the scene. According to the conversational maxims of Grice (cf. Grice [1975]), the listener should be informed about all relevant facts and redundancy should be avoided.

Relevance depends on factors like: (i) salience, which is determined by the frequency of occurrence and the complexity of the generic event or action concept, (ii) topicality, and (iii) current state, i.e., fully recognized occurrences are preferred. Topicality decreases for terminated movements and actions as the scene progresses and during recognition events and plans enter different states, i.e., relevance changes continually. To avoid redundancy, an occurrence will not be mentioned if it is implied by some other proposition already verbalized, e.g., a `have-ball` event following a `pass` will not be selected for verbalization.

Additional selection processes are used to determine deep cases and to choose descriptions for objects, locations, and time; in these choices the contents of the text memory and the listener model must also be considered.

The linearization process determines the order in which the selected propositions should be mentioned in the text. The temporal ordering of the corresponding events and actions is the primary consideration for linearization; secondarily, focusing criteria are used to maintain discourse coherence.

## 4.2 Anticipating the listener's visual imagination

After relevant propositions are selected and ordered, they are passed on to the listener model ANTLIMA (cf. Schirra and Stopp [1993]), which constructs a “*mental image*” corresponding to the visual conceptualizations that the system's utterance would elicit in the listener's mind. The (assumed) imagination is compared with the system's visual information and incompatibilities are fed back to the generation component in order to adjust the preliminary text plan. A similar *anticipation feedback loop*, has been proposed in (cf. Jameson and Wahlster [1982]) for the generation of pronouns.

A plausible mental image is constructed by searching for a maximally typical representation of a situation described by the selected propositions. The typicality distribution corresponding to a certain proposition is encoded in a so-called *Typicality Potential Field* (TyPoF), a function mapping locations to typicality values. TyPoFs are instances of typicality schemas associated with spatial relations as well as event and action concepts. Each TyPoF takes into account the dimensionality, size, and shape of the objects involved. In Fig. 5, the TyPoFs for *'player A in front of player B'* and *'in front of the goal area'* are visualized. A typicality value associated with a spatial expression corresponds to the (degree of) applicability of a spatial relation for a given object configuration.

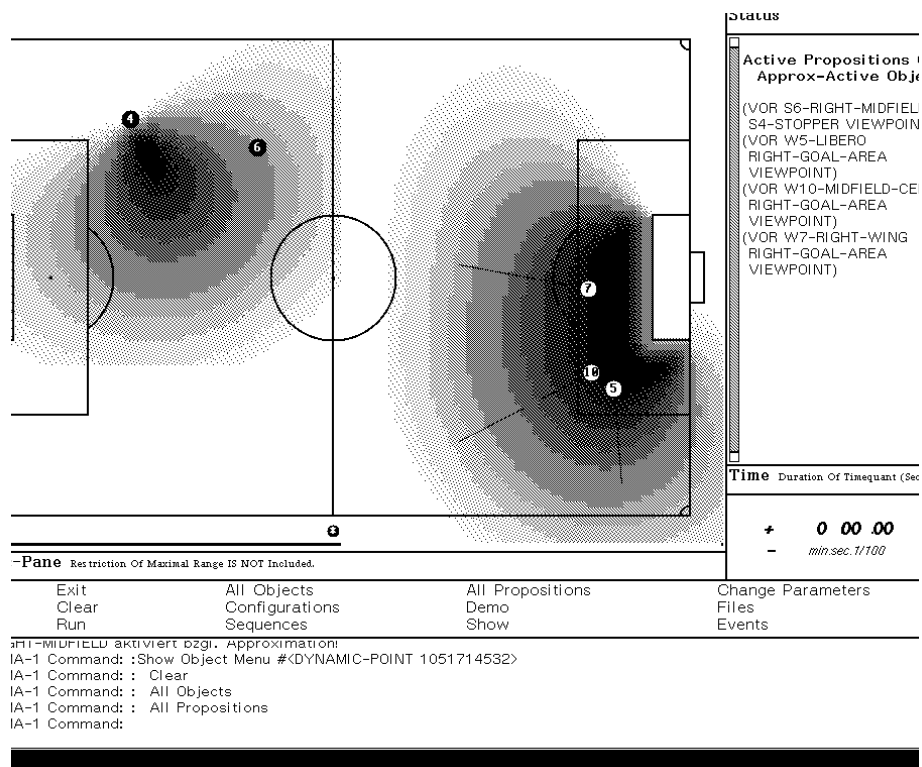


Figure 5: Examples of typicality distributions

If several propositions impose restrictions on an object, the corresponding TyPoFs are combined by taking the average. In the case of incompatible restrictions the preliminary text plan has to be retracted. Hillclimbing is employed in order to find an interpretation with maximal typicality. Then the mental image is re-analyzed, i.e., the processes of high-level analysis are applied to it. The resulting set of propositions is compared to the propositions computed from the image sequence and detected misunderstandings may be dispelled by changing the preliminary text plan.

### 4.3 Incremental verbalization

The encoding of the selected propositions includes lexicalization, the determination of morphosyntactic information, and surface transformations.

In the process of transforming symbolic event descriptions into natural language utterances, first a verb is selected by accessing the concept lexicon, and the case-roles associated with the verb are instantiated. Control passes back to the selection component, which decides which information concerning the case-role fillers should be conveyed. The selected information is transformed into natural-language expressions referring to time, space or objects. Time is indicated by the verb tense and by temporal adverbs; spatial prepositions and appropriate objects of reference are selected to refer

to spatial relations. Internal object identifiers are transformed into noun phrases by the selection of attributes that enable the listener to uniquely identify the intended referent. If an object cannot be characterized by attributes stored *a priori* in the partner model, it will be described by means of spatial relations, such as *'the left goal'*, or by means of occurrences already mentioned in which it was (is) involved, e.g., *'the player who was attacked'*. Anaphoric expressions are generated if the referent is in focus and no ambiguity is possible.

Recognized intentions can be reflected in natural language descriptions in various ways. For instance, they can be expressed explicitly (*'She wants to do A'*) or be construed as expectations and formulated in the future tense. They can also be expressed implicitly, using verbs that imply intention (e.g., *'chase'*). In addition, relationships between intentions and actions or among several intentions of a single agent can be described, e.g., using purposive clauses (*'He did A in order to achieve B'*). Cooperative interactions can be summarized most easily, using a natural language expression describing the collective intention. Cooperative as well as antagonistic interactions can be described in more detail using temporal adverbs and conjunctions. Plan failures can also be stated explicitly, or they can be related to their causes by means of causal clauses. In our current implementation it is only possible to explicitly express intentions and relationships between intentions of a single agent.

To meet the requirements of simultaneous scene description, information concerning partly-recognized events and actions is also provided. Consequently, language generation cannot start from completely worked-out conceptual contents; i.e., the need for an incremental generation strategy arises (see, e.g., Reithinger [1992]). In the newest version of the VITRA system the incremental generation of surface structures is realized with the module described in (cf. Harbusch et al. [1991], Finkler and Schauder [1992]), an incremental generator for German and English, which is based on *Tree Adjoining Grammars*.

## 5 Conclusion

VITRA is the first system that automatically generates natural language descriptions for recognized trajectories of objects in a real world image sequence. High-level scene analysis in VITRA is not restricted to the purely visual, i.e., spatio-temporal, properties of the scene, but also aims at the recognition of presumed goals and plans of the observed agents. In addition, the listener model in VITRA anticipates the (assumed) imagination of the listener for the generation of the most appropriate description.

Our approach towards *simultaneous* scene description emphasizes concurrent image sequence evaluation and natural language processing. The processing in all sub-components is carried out on an *incremental* basis, and hence provides an important prerequisite for real-time performance.

Despite these promising results, we are still far away from a universally applicable AI system capable of describing an arbitrary sequence of images. Nonetheless, the

VITRA system will serve as a workbench for the further investigation of problems arising in the field of integrated vision and natural language processing.

In order to improve the quality of text production in the VITRA prototype, the language generation component will be extended for the description of plan failures and interactions, i.e., information that can already be provided by the high-level scene analysis.

So far, we have only been concerned with a bottom-up analysis of image sequences, recorded with a stationary TV-camera. Future work will concentrate on expectation-driven scene analysis. Intermediate results of the high-level analysis shall support low-level vision in focussing on relevant objects and in providing parameters for the active control of the sensor adjustment. On the one hand, focussing techniques are necessary to compensate the computational complexity of the analysis in more advanced applications, on the other hand, interaction between low-level and high-level analysis is required if VITRA is to become robust for the difficulties caused by insufficient low-level image processing. These issues will be studied in the context of natural language interaction with an autonomous mobile robot, equipped with several sensors.

## 6 Technical Notes

The current version of the VITRA system is written in Common Lisp and CLOS, with the graphical user interface implemented in CLIM. The system has been developed on Symbolics 36xx Lisp Machines, Symbolics UX1200S Lisp Coprocessors, and on Hewlett Packard 9720 and SPARC Workstations.

## Acknowledgements

The work described here was partly supported by the Sonderforschungsbereich 314 der Deutschen Forschungsgemeinschaft, “Künstliche Intelligenz und wissensbasierte Systeme” Projekt N2: VITRA.

We would like to thank Paul Mc Kevitt and an anonymous reviewer for their helpful comments on an earlier version of this article.

## References

- E. André, G. Bosch, G. Herzog, T. Rist.** Coping with the Intrinsic and the Deictic Uses of Spatial Prepositions. In: K. Jorrand, L. Sgurev, eds., *Artificial Intelligence II: Methodology, Systems, Applications*, pp. 375–382, North-Holland, Amsterdam, 1987a.

- E. André, G. Herzog, T. Rist.** On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER. In: *Proc. of the 8th ECAI*, pp. 449–454, Munich, 1988.
- E. André, G. Herzog, T. Rist.** Natural Language Access to Visual Data: Dealing with Space and Movement. Report 63, Universität des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1989, Presented at the 1st Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France.
- E. André, T. Rist, G. Herzog.** Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung zeitveränderlicher Szenen. In: K. Morik, ed., *GWAI-87. 11th German Workshop on Artificial Intelligence*, pp. 330–337, Springer, Berlin, Heidelberg, 1987b.
- R. Bajcsy, A. K. Joshi, E. Krotkov, A. Zwarico.** LandScan: A Natural Language and Computer Vision System for Analyzing Aerial Images. In: *Proc. of the 9th IJCAI*, pp. 919–921, Los Angeles, CA, 1985.
- W. Finkler, A. Schauder.** Effects of Incremental Output on Incremental Natural Language Generation. In: *Proc. of the 10th ECAI*, pp. 505–507, Vienna, 1992.
- K.-P. Gapp.** Berechnungsverfahren für räumliche Relationen in 3D-Szenen. Memo 59, Universität des Saarlandes, SFB 314, 1993.
- K.-P. Gapp.** Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space. In: *Proc. of AAAI-94*, pp. 1393–1398, Seattle, WA, 1994.
- H. P. Grice.** Logic and Conversation. In: P. Cole, J. L. Morgan, eds., *Speech Acts*, pp. 41–58, Academic Press, London, 1975.
- K. Harbusch, W. Finkler, A. Schauder.** Incremental Syntax Generation with Tree Adjoining Grammars. In: W. Brauer, D. Hernández, eds., *Verteilte Künstliche Intelligenz und kooperatives Arbeiten: 4. Int. GI-Kongreß Wissensbasierte Systeme*, pp. 363–374, Springer, Berlin, Heidelberg, 1991.
- G. Herzog.** Ein Werkzeug zur Visualisierung und Generierung von geometrischen Bildfolgenbeschreibungen. Memo 12, Universität des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1986.
- G. Herzog.** Utilizing Interval-Based Event Representations for Incremental High-Level Scene Analysis. In: M. Aurnague, A. Borillo, M. Borillo, M. Bras, eds., *Proc. of the 4th International Workshop on Semantics of Time, Space, and Movement and Spatio-Temporal Reasoning*, pp. 425–435, Château de Bonas, France, Groupe “Langue, Raisonnement, Calcul”, Toulouse, 1992a.
- G. Herzog.** Visualization Methods for the VITRA Workbench. Memo 53, Universität des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1992b.

- G. Herzog, W. Maaß, P. Wazinski.** VITRA GUIDE: Utilisation du Langage Naturel et de Représentation Graphiques pour la Description d'Itinéraires. In: *Images et Langages: Multimodalité et Modélisation Cognitive, Colloque Interdisciplinaire du Comité National de la Recherche Scientifique*, pp. 243–251, Paris, 1993a.
- G. Herzog, J. Schirra, P. Wazinski.** Arbeitsbericht für den Zeitraum 1991–1993: VITRA – Kopplung bildverstehender und sprachverstehender Systeme. Memo 58, Universität des Saarlandes, SFB 314 (VITRA), Saarbrücken, 1993b.
- G. Herzog, C.-K. Sung, E. André, W. Enkelmann, H.-H. Nagel, T. Rist, W. Wahlster, G. Zimmermann.** Incremental Natural Language Description of Dynamic Imagery. In: C. Freksa, W. Brauer, eds., *Wissensbasierte Systeme. 3. Int. GI-Kongreß*, pp. 153–162, Springer, Berlin, Heidelberg, 1989.
- A. Jameson, W. Wahlster.** User Modelling in Anaphora Generation. In: *Proc. of the 5th ECAI*, pp. 222–227, Orsay, France, 1982.
- D. Koller.** *Detektion, Verfolgung und Klassifikation bewegter Objekte in monokularen Bildfolgen am Beispiel von Straßenverkehrsszenen.* Infix, St. Augustin, 1992.
- D. Koller, K. Daniilidis, T. Thórhallson, H.-H. Nagel.** Model-based Object Tracking in Traffic Scenes. In: G. Sandini, ed., *Proc. of Second European Conf. on Computer Vision*, pp. 437–452, Springer, Berlin, Heidelberg, 1992a.
- D. Koller, N. Heinze, H.-H. Nagel.** Algorithmic Characterization of Vehicle Trajectories from Image Sequences by Motion Verbs. In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 90–95, Maui, Hawaii, 1992b.
- H. Kollnig, H.-H. Nagel.** Ermittlung von begrifflichen Beschreibungen von Geschehen in Straßenverkehrsszenen mit Hilfe unscharfer Mengen. *Informatik Forschung und Entwicklung*, **8**(4), 186–196, 1993.
- T. C. Lüth, T. Längle, G. Herzog, E. Stopp, U. Rembold.** KANTRA: Human-Machine Interaction for Intelligent Robots Using Natural Language. In: *3rd IEEE Int. Workshop on Robot and Human Communication, RO-MAN'94*, pp. 106–111, Nagoya, Japan, 1994.
- W. Maaß, P. Wazinski, G. Herzog.** VITRA GUIDE: Multimodal Route Descriptions for Computer Assisted Vehicle Navigation. In: *Proc. of the Sixth Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE-93*, pp. 144–147, Edinburgh, Scotland, 1993.
- B. Neumann.** Natural Language Description of Time-Varying Scenes. In: D. L. Waltz, ed., *Semantic Structures: Advances in Natural Language Processing*, pp. 167–207, Lawrence Erlbaum, Hillsdale, NJ, 1989, ISBN 0-89859-817-6.

- B. Neumann, H.-J. Novak.** NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen. *Informatik Forschung und Entwicklung*, **1**, 83–92, 1986.
- H. Niemann, H. Bunke, I. Hofmann, G. Sagerer, F. Wolf, H. Feistel.** A Knowledge Based System for Analysis of Gated Blood Pool Studies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **7**, 246–259, 1985.
- N. Reithinger.** The Performance of an Incremental Generation Component for Multi-Modal Dialog Contributions. In: R. Dale, E. Hovy, D. Rösner, O. Stock, eds., *Aspects of Automated Natural Language Generation: Proc. of the 6th Int. Workshop on Natural Language Generation*, pp. 263–276, Springer, Berlin, Heidelberg, 1992.
- G. Retz-Schmidt.** Various Views on Spatial Prepositions. *AI Magazine*, **9**(2), 95–105, 1988.
- G. Retz-Schmidt.** Recognizing Intentions, Interactions, and Causes of Plan Failures. *User Modeling and User-Adapted Interaction*, **1**, 173–202, 1991.
- G. Retz-Schmidt.** *Die Interpretation des Verhaltens mehrerer Akteure in Szenenfolgen*. Springer, Berlin, Heidelberg, 1992.
- K. Rohr.** Towards Model-based Recognition of Human Movements in Image Sequences. *Computer Vision, Graphics, and Image Processing (CVGIP): Image Understanding*, **59**(1), 94–115, 1994.
- J. R. J. Schirra, G. Bosch, C.-K. Sung, G. Zimmermann.** From Image Sequences to Natural Language: A First Step Towards Automatic Perception and Description of Motions. *Applied Artificial Intelligence*, **1**, 287–305, 1987.
- J. R. J. Schirra, E. Stopp.** ANTLIMA—A Listener Model with Mental Images. In: *Proc. of the 13th IJCAI*, pp. 175–180, Chambéry, France, 1993.
- C.-K. Sung.** Extraktion von typischen und komplexen Vorgängen aus einer langen Bildfolge einer Verkehrsszene. In: H. Bunke, O. Kübler, P. Stucki, eds., *Mustererkennung 1988; 10. DAGM Symposium*, pp. 90–96, Springer, Berlin, Heidelberg, 1988.
- C.-K. Sung, G. Zimmermann.** Detektion und Verfolgung mehrerer Objekte in Bildfolgen. In: G. Hartmann, ed., *Mustererkennung 1986; 8. DAGM-Symposium*, pp. 181–184, Springer, Berlin, Heidelberg, 1986.
- J. K. Tsotsos.** Knowledge Organization and its Role in Representation and Interpretation for Time-Varying Data: the ALVEN System. *Computational Intelligence*, **1**, 16–32, 1985.

- W. Wahlster.** One Word Says More Than a Thousand Pictures. On the Automatic Verbalization of the Results of Image Sequence Analysis Systems. *Computers and Artificial Intelligence*, **8**(5), 479–492, 1989.
- W. Wahlster, H. Marburger, A. Jameson, S. Busemann.** Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. In: *Proc. of the 8th IJCAI*, pp. 643–646, Karlsruhe, FRG, 1983.
- I. Walter, P. C. Lockemann, H.-H. Nagel.** Database Support for Knowledge-Based Image Evaluation. In: P. M. Stocker, W. Kent, R. Hammersley, eds., *Proc. of the 13th Conf. on Very Large Databases, Brighton, UK*, pp. 3–11, Morgan Kaufmann, Los Altos, CA, 1988.
- P. Wazinski.** Graduated Topological Relations. Memo 54, Universität des Saarlandes, SFB 314, 1993a.
- P. Wazinski.** Graduierte topologische Relationen. In: D. Hernandez, ed., *Hybride und integrierte Ansätze zur Raumrepräsentation und ihre Anwendung, Workshop auf der 17. Fachtagung für Künstliche Intelligenz, Berlin*, pp. 16–19, Technische Univ. München, Institut für Informatik, 1993b, Forschungsberichte Künstliche Intelligenz, FKI-185-93.